Low-Resource Dialogue Summarization with Domain-Agnostic Multi-Source Pretraining

Yicheng Zou[1,2], **Bolin Zhu**[2], **Xingwu Hu**[2], **Tao Gui**[1*], **Qi Zhang**[2*]
[1]Institute of Modern Languages and Linguistics, Fudan University
[2]Shanghai Key Laboratory of Intelligent Information Processing, Fudan University
[2]School of Computer Science, Fudan University
Shanghai, China
{yczou18,blzhu20,xwhu20,tgui,qz}@fudan.edu.cn

Abstract

With the rapid increase in the volume of dialogue data from daily life, there is a growing demand for dialogue summarization. Unfortunately, training a large summarization model is generally infeasible due to the inadequacy of dialogue data with annotated summaries. Most existing works for low-resource dialogue summarization directly pretrain models in other domains, e.g., the news domain, but they generally neglect the huge difference between dialogues and conventional articles. To bridge the gap between out-of-domain pretraining and in-domain fine-tuning, in this work, we propose a multi-source pretraining paradigm to better leverage the external summary data. Specifically, we exploit large-scale in-domain non-summary data to separately pretrain the dialogue encoder and the summary decoder. The combined encoder-decoder model is then pretrained on the out-of-domain summary data using adversarial critics, aiming to facilitate domain-agnostic summarization. The experimental results on two public datasets show that with only limited training data, our approach achieves competitive performance and generalizes well in different dialogue scenarios.

1 Introduction

With the explosion in the quantity of dialogue data from the Internet and daily life, there is growing interest in automatic dialogue summarization for various scenarios and applications, such as email threads, meetings, customer service, and online chats (Murray and Carenini, 2008; Shang et al., 2018; Liu et al., 2019; Zou et al., 2021a,b). Unfortunately, creating large-scale dialogue datasets with annotated summaries is costly and labor-intensive, which makes it difficult to build and train large summarization models using adequate supervision signals, especially in a new domain. Hence, it is necessary to develop models for dialogue summarization in low-resource settings, where only limited or even no training examples are available.

Recently, domain adaptation approaches with large-scale pretraining have attracted much attention in low-resource summarization (Wang et al., 2019; Yang et al., 2020; Zhang et al., 2020). A similar strategy is used in dialogues, whereby external summary data from other domains, e.g., the CNN/Dailymail news dataset (Hermann et al., 2015), are introduced for model pretraining prior to the final fine-tuning on low-resource dialogue summaries. Recent works (Gliwa et al., 2019; Zhu et al., 2020; Joshi et al., 2020) have also reported the effectiveness of pretrained summarizers for different kinds of dialogue scenarios, such as chat logs and medical conversations.

However, dialogue summary data has several inherent and significant differences from conventional articles in terms of text styles and summary structures. (i) Dialogues generally contain multiple participants who have distinct characteristics. (ii) Rather than the formal expressions found in news documents, dialogues often comprise utterances with informal or ungrammatical phrases. (iii) The structure of a dialogue summary, including length and the level of abstraction, is quite different from that in other domains (Zhu et al., 2020), e.g., CNN/Dailymail. Thus, considering the huge difference between dialogues and general documents, direct finetuning on dialogue summaries is not ideal when using a model pretrained from other domains.

To better leverage summary data from domains such as news or scientific articles, in this work, we introduce a novel pretraining paradigm called domain-agnostic multi-source pretraining (DAMS) to summarize dialogues in a low-resource setting. We postulate that the pretraining of dialogue summarization could be decomposed into three procedures: the pretraining of encoder, decoder, and the combined encoder-decoder model. Specifically, the dialogue encoder is pretrained on large-

[*] Corresponding authors.

scale unannotated dialogues to learn the way of dialogue modeling and understanding. The summary decoder is pretrained on large-scale summary-like short texts to learn a language model in the style of the dialogue summaries. Furthermore, the encoder and decoder are combined and pretrained on external summary data to go through an integral process of summarization. The above pretraining processes from the three sources are performed simultaneously. By this means, DAMS exploits large-scale non-summary data in the same domain to narrow the gap between pretraining and fine-tuning. Additionally, adversarial critics are used to capture the features shared between dialogues and general documents, and to learn to perform domain-agnostic summarization.

We conducted experiments on two public dialogue summary datasets, namely SAMSum (Gliwa et al., 2019) and ADSC (Misra et al., 2015). Pretraining was conducted on datasets from multiple sources, including dialogue corpora, daily-life short text corpora, and text summarization datasets from the news domain. The experimental results show that with only limited training data of dialogue summaries, our approach achieved competitive performance and showed a promising ability for generalizing different dialogue scenarios. Our codes and datasets are publicly available[1].

In summary, our contributions are three-fold: 1) We explore the task of dialogue summarization in a low-resource setting with the usage of external multi-source corpora. 2) A novel pretraining strategy is designed to bridge the gap between out-of-domain pretraining and in-domain fine-tuning for domain-agnostic summarization. 3) Comprehensive studies on two datasets show the effectiveness of our method in various aspects.

2 Related Work

2.1 Dialogue Summarization

Dialogue summarization is a challenging and valuable task that receives much attention in recent years. Different from studies on conventional documents like news or reviews (See et al., 2017; Narayan et al., 2018; Chu and Liu, 2019), dialogue summarization is investigated in multi-party interactions such as mail threads (Rambow et al., 2004), meetings (Gillick et al., 2009; Shang et al., 2018; Zhong et al., 2021), telephone conversation records (Zechner, 2001; Gurevych and Strube, 2004), and daily chats (Gliwa et al., 2019; Zhao et al., 2020). Most of these approaches share a similar prerequisite: a decent labeled training dataset with annotated summaries. Nevertheless, creating a large-scale dialogue summary dataset is very expensive and labor-intensive, which makes the traditional methods hard to apply in real-world applications, especially when only limited or even no training signals are available. In this work, we explore dialogue summarization in a low-resource setting, and leverage external large-scale corpora to facilitate the task, which is applicable to most dialogue scenarios.

2.2 Domain Adaptation for Summarization

Since texts and their summaries across diverse domains might share similarities and benefit from each other, domain adaptation for text summarization has attracted much research interest recently (Hua and Wang, 2017; Wang et al., 2019; Zhang et al., 2020; Yang et al., 2020; Yu et al., 2021). Most existing works perform pretraining on large-scale out-of-domain datasets and then adapt to the in-domain summary data. For dialogue summarization, although it is more ideal to perform adaptation from a source dialogue domain to a target dialogue domain (Sandu et al., 2010; Wang and Cardie, 2013), unfortunately, the inadequacy of dialogue summary data makes it infeasible to directly train a large summarization model on the source data in an end-to-end manner. Recently, a couple of works have leveraged large-scale summary data that is more distinct from the dialogue domain, e.g., the news domain, to facilitate dialogue summarization (Gliwa et al., 2019; Zhu et al., 2020; Joshi et al., 2020). However, the huge gap between dialogues and general articles is barely noticed. Yu et al. (2021) conducted pretraining on the news summary data and the dialogue non-summary data simultaneously, but the two different tasks share a single decoder, which might confuse the model about the knowledge that it learns. To better leverage the out-of-domain summary data and the in-domain non-summary data, we explore the domain-agnostic summarization. It is supported by a multi-source pretraining paradigm with adversarial learning, where the encoder and the decoder are separately pretrained on the in-domain non-summary data and combinedly pretrained on the out-of-domain summary data, aiming to narrow the gap between pretraining and fine-tuning.

[1] https://github.com/RowitZou/DAMS

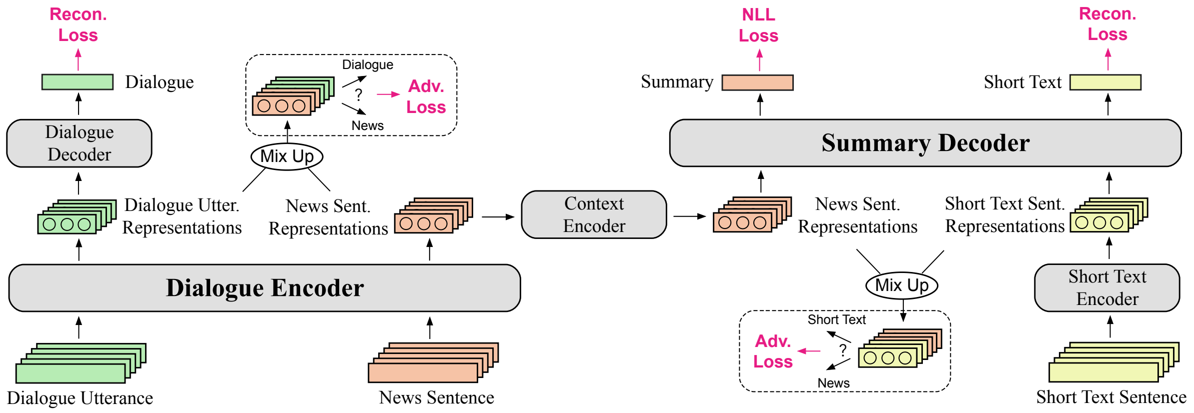


Figure 1: The overall architecture of DAMS. The multi-source pretraining includes: (i) encoder pretraining using dialogues (green); (ii) decoder pretraining using short texts (yellow); (iii) Joint pretraining using general articles with corresponding summaries (orange).

3 Methodology

In this section, we detail the low-resource dialogue summarization under the domain-agnostic multi-source pretraining (DAMS). It consists of three pretraining objectives: two reconstruction losses with denoising auto-encoders that learn dialogue modeling and summary-like text generation; a sequence-to-sequence (seq2seq) training objective with the combined dialogue encoder and summary decoder that learns abstractive summarization. Additionally, two adversarial critics are attached to the encoder's output representations and the decoder's input representations, learning to perform domain-agnostic summarization. The overall framework is illustrated in Figure 1.

3.1 Multi-Source Pretraining

Despite the considerable amount of summary data in other domains such as news and scientific articles, adaptation to the dialogue domain is not easy due to the huge difference between dialogues and conventional articles. To address this issue, we postulate that abstractive dialogue summarization could be decomposed into three procedures: (i) Dialogue modeling for understanding dialogue semantics and capturing dialogue characteristics; (ii) Saliency estimation based on learned representations to identify the important parts of input contents; (iii) Generating a summary grounded on the salient information with a certain style or structure. Although the limited dialogue summary data is inadequate to train the three procedures jointly, each one of them, fortunately, could be well handled by separate pretraining with large-scale corpora from different sources. Specifically, dialogue modeling can benefit from the usage of large-scale unannotated dialogues. The external news summary data may contribute to the process of saliency estimation. A language model trained on daily-life short texts can generate discourses with the style of dialogue summaries, rather than formal expressions in news or scientific articles.

Pretraining of dialogue modeling. Inspired by recent large-scale pretraining models (Devlin et al., 2019; Zhang et al., 2019; Lewis et al., 2020), we exploit the framework of denoising auto-encoding (DAE) (Vincent et al., 2008) to extract robust features to compose dialogue representations. Formally, we denote each dialogue as an utterance sequence $D = \{u_1, u_2 ..., u_n\}$. To incorporate multi-party information, we add the name of the speaker at the beginning of each utterance. Then, we tokenize utterances into word sequences, denoted as $u_i = \{w_{i1}, .., w_{im}\}$, where w_{ij} is the j-th word in the sequence of u_i. For noise addition, we randomly mask 15% of the tokens in each utterance with a special [MASK] token similar to BERT[2] (Devlin et al., 2019). The purpose of noise addition is to encourage DAE to reconstruct the original utterances for robust representation learning.

In this work, we employ Transformer with multi-head attentions (Vaswani et al., 2017) as the basic encoder and decoder of DAE. Before inputting word sequences into the encoder, we concatenate a special token [CLS] in front of each sequence similar to BERT. The final hidden state

[2] In practise, we keep 20% of the utterances unchanged. The purpose of this is to bias the representation towards the actual observed utterance.

corresponding to this token is used as the aggregate sequence representation for utterance reconstruction. Formally, we transform the modified noisy sequence $\widetilde{u}_i = \{w_i^{cls}, w_{i1}..., w_{im}\}$ into a sequence of hidden vectors by a Transformer encoder:

$$[\mathbf{h}_i^{cls}, \mathbf{h}_{i1}, ..., \mathbf{h}_{im}] = \mathrm{TF}_{\theta_e^d}([\mathbf{e}_i^{cls}, \mathbf{e}_{i1}, ..., \mathbf{e}_{im}]), \tag{1}$$

where $\mathbf{e}_{ij}$ is the embedding of the j-th word w_{ij} in the word sequence, while $w_i^{cls}, \mathbf{e}_i^{cls}$ represent [CLS] and its embedding.

The decoder is an auto-regressive model that recovers the original utterance conditioned on the input representation $\mathbf{h}_i^{cls}$. Here, we use a Transformer decoder with masked attention that conditions by adding $\mathbf{h}_i^{cls}$ to each input embedding. This is a Transformer variant that removes the decoder-encoder attention layer. Formally, the generation probability is defined by:

$$P(\hat{w}_{ij}|\hat{w}_{i(1:j-1)}; \widetilde{u}_i) = \mathrm{TF}_{\theta_g^d}([\hat{\mathbf{e}}_{i(1:j-1)}]; \mathbf{h}_i^{cls}), \tag{2}$$

where $\hat{\mathbf{e}}_{ij}$ denotes the embedding of the predicted word $\hat{w}_{ij}$ at the decoding step j. Notably, the decoder applies utterance representations $\mathbf{h}_i^{cls}$ as memories instead of using word-level attention or copy mechanism. It encourages all semantics to be captured in $\mathbf{h}_i^{cls}$. In Section 3.4, we give a further discussion about why we do not choose the word-level cross-attention. Finally, we use the original utterance u_i as a gold reference to train the DAE for utterance reconstruction on large-scale dialogue corpora, paving the way to dialogue modeling for the downstream summarization task:

$$\mathcal{L}_{rec} = -\sum\nolimits_i \sum\nolimits_{j=1}^{m} \log P(\hat{w}_{ij}|\hat{w}_{i(1:j-1)}; \widetilde{u}_i). \tag{3}$$

Pretraining of summary language modeling. We use the similar strategy as in dialogue pretraining to learn a summary language model. Here, we introduce the external corpora that contain daily-life short texts or stories, e.g., BooksCorpus (Zhu et al., 2015), to train the decoder to generate texts in the style of dialogue summaries. We truncate long documents into text pieces to form training samples, each one of which includes several consecutive sentences. We also add noise to these text pieces and train a DAE to recover them. Specifically, given the sentence sequence of a training sample $S = \{s_1, s_2, ..., s_n\}$, we use the same noise addition strategy as for dialogues to construct noisy sentences, and encode them into hidden vectors by a Transformer encoder $\mathrm{TF}_{\theta_e^s}$ similar to Eq.1.

The generation process, however, is different from that of utterance reconstruction. Since a summary might contain more than one sentence, we should encourage the decoder to generate all sentences of S sequentially to simulate the process of summary generation. Hence, to further capture the global semantic dependency between sentences, we use another Transformer encoder to hierarchically fuse context information:

$$[\hat{\mathbf{h}}_1^{cls}, \hat{\mathbf{h}}_2^{cls}, ..., \hat{\mathbf{h}}_n^{cls}] = \mathrm{TF}_{\theta_h^s}([\mathbf{h}_1^{cls}, \mathbf{h}_2^{cls}, ..., \mathbf{h}_n^{cls}]). \tag{4}$$

Here, all sentence representations derived from [CLS] tokens are fed into the hierarchical encoder for information interaction. The output vectors are then used as memories for decoder-encoder attention in a classic Transformer decoder to recover S. The generation probability is:

$$\begin{aligned} \hat{\mathbf{H}}^{cls} &= [\hat{\mathbf{h}}_1^{cls}, \hat{\mathbf{h}}_2^{cls}, ..., \hat{\mathbf{h}}_n^{cls}], \\ P(\hat{w}_k|\hat{w}_{1:k-1}; \widetilde{S}) &= \mathrm{TF}_{\theta_g^s}([\hat{\mathbf{e}}_{1:k-1}]; \hat{\mathbf{H}}^{cls}), \end{aligned} \tag{5}$$

where $\widetilde{S}$ represents the noisy text piece and $\hat{w}_k, \hat{\mathbf{e}}_k$ denote the k-th predicted word and its embedding. The difference between Eq.2 and Eq.5 is that the former reconstructs a single utterance, while the latter predicts the entire text sample. Finally, we train the language model conditioned on $\widetilde{S}$ as:

$$\mathcal{L}_{gen} = -\sum\nolimits_k \log P(\hat{w}_k|\hat{w}_{1:k-1}; \widetilde{S}). \tag{6}$$

Pretraining of abstractive summarization. In order to pretrain end-to-end summary generation, we bridge the dialogue encoder $\mathrm{TF}_{\theta_e^d}$ with the summary decoder $\mathrm{TF}_{\theta_g^s}$ using a context encoder $\mathrm{TF}_{\theta_h^b}$. $\mathrm{TF}_{\theta_h^b}$ has the same architecture as in Eq.4. Then, we input sentences of a document into $\mathrm{TF}_{\theta_e^d}$ and get a predicted summary from $\mathrm{TF}_{\theta_g^s}$, training the model with the following objective:

$$\mathcal{L}_{summ} = -\sum\nolimits_k \log P(\hat{w}_k|\hat{w}_{1:k-1}; D_s), \tag{7}$$

where D_s is the document. $\hat{w}_k$ represents the k-th word in the predicted summary. Here, we reuse $\mathrm{TF}_{\theta_e^d}$ and $\mathrm{TF}_{\theta_g^s}$ for abstractive summarization, and its purpose is to bridge the gap between separate pretraining on multi-source texts and joint fine-tuning on dialogue summaries. By an integral process of text summarization, the combined encoder-decoder model learns to capture salient information from sentence (or utterance) representations and generate summaries accordingly.

3.2 Domain-Agnostic Summarization with Adversarial Learning

Ideally, the DAE learns a high-level latent *content* conveyed in representations, disentangled from their original attributes, e.g., styles of informal dialogue utterances and formal news sentences, adapting the way of saliency estimation and summary generation to the dialogue domain. However, models often learn domain-specific features, making it difficult to generalize in a new domain (Peng et al., 2019). To address this issue, inspired by recent works of adversarial summary generation (Liu et al., 2018; Rekabdar et al., 2019), we add an adversarial discriminator (critic) that learns to identify the domain of each representation, and use a gradient reversal mechanism (Ganin and Lempitsky, 2015) to ensure that the feature distributions over different domains are made similar (as indistinguishable as possible for the discriminator), thus resulting in the domain-invariant features and encouraging the summarizer to only focus on content rather than domain-specific attributes.

Here, we add two adversarial critics $\mathrm{D}_e, \mathrm{D}_g$ on the output vectors of $\mathrm{TF}_{\theta_e^d}$ and the input vectors of $\mathrm{TF}_{\theta_g^s}$, respectively (see Figure 1). The former classifies output vectors as dialogue utterances or news sentences, and the latter tries to distinguish news articles from short texts. The adversarial critic is a simple binary classifier with a multi-layer perceptron and a sigmoid activator trained by a logistic loss function, denoted as $\mathcal{L}_e^D, \mathcal{L}_g^D$ for D_e and D_g, respectively. Finally, we combine all pretraining losses and adversarial signals to jointly train the model, where α is a hyper-parameter to adjust the loss proportion:

$$\mathcal{L} = \mathcal{L}_{rec} + \mathcal{L}_{gen} + \mathcal{L}_{summ} + \alpha(\mathcal{L}_e^D + \mathcal{L}_g^D). \quad (8)$$

3.3 Fine-tuning on Dialogue Summaries

After multi-source pretraining, we further stack $\mathrm{TF}_{\theta_e^d}$, $\mathrm{TF}_{\theta_h^b}$, and $\mathrm{TF}_{\theta_g^s}$ for joint fine-tuning on the dialogue summary dataset. The learning objective is similar to Eq.7. Notably, the three modules are fully trained by appropriate data from multiple sources, leading to a higher convergence speed on the target dialogue summaries (see details in Section 5.3), which requires fewer training data points to achieve a competitive performance.

Dataset	Split	# of dial.	Avg. words	Avg. turns	Ref. length
SAMSum	Train	14,732	120.26	11.13	22.81
	Dev.	818	117.46	10.72	22.80
	Test	819	122.71	11.24	22.47
ADSC	All	45	370.44	7.51	101.99

Table 1: Statistics of dialogue summary datasets.

3.4 Discussion of the Encoder-Decoder Connection Strategy

The encoder-decoder cross attention for encoding the context information is widely used in transformer-based architectures. Large-scale pretraining models for the summarization task, e.g., BART (Lewis et al., 2020), generally exploit token-level attention to integrate the document context. In this work, we have tried keeping the traditional token-level cross attention in the proposed architecture to directly connect the dialogue encoder and the summary decoder. However, we find that it is difficult to disentangle the encoder and the decoder for separate pretraining. It is also hard to add adversarial critics to token-level representations involved in the cross attention to learn domain-invariant features. Considering the above limitations, we use an embedding concatenation strategy in the dialogue decoder $\mathrm{TF}_{\theta_g^d}$ as a DAE to learn utterance representations. The summary decoder $\mathrm{TF}_{\theta_g^s}$ still has the cross attention, but keys and values are sentence representations from the context encoder $\mathrm{TF}_{\theta_h^b}$ instead of token representations from the dialogue encoder $\mathrm{TF}_{\theta_e^d}$. Here, $\mathrm{TF}_{\theta_h^b}$ bridges the dialogue encoder and the summary decoder. It not only captures the context information of sentences (utterances), but also derives sentence-level representations that are applicable for domain identification in adversarial learning. Nevertheless, the abandonment of token-level attention will inevitably affect the fine-grained information integration. In terms of how to keep the token-level cross attention in DAMS, we leave it as a future work for open discussions.

4 Experimental Settings

4.1 Datasets

Following the latest works (Zhao et al., 2020; Feng et al., 2020), we evaluate our method on two public dialogue summary datasets SAMSum

(Gliwa et al., 2019) and ADSC[3] (Misra et al., 2015). Statistics of the dialogue datasets is shown in Table 1. SAMSum originally contains 14k training examples. To simulate a low-resource scenario, we start from using the full training data, and gradually reduce the number of training examples by halving the training set. For multi-source pretraining, we use the following datasets.

Dialogues. We use Reddit Conversation Corpus (Dziri et al., 2019)[4] for the pretaining of dialogue modeling. It contains about 15M context-response pairs for training, where each dialogue context consists of 3.5 utterances on average.

Short Texts. We choose MSCOCO (Lin et al., 2014) and BookCorpus (Zhu et al., 2015) to pretrain the summary language model. MSCOCO is a standard benchmark dataset for the image caption generation task, which contains over 120K images and 600K captions describing the prominent object/action in an image. Here, we only use captions to train the generator. BookCorpus is a large-scale corpus containing 11,038 free books from the Internet. We randomly truncate long documents into text pieces as training samples[5]. Each sample contains 1.5 sentences on average and we collect about 5M samples for training.

Summarization Corpus. CNN/DailyMail (Hermann et al., 2015), Gigaword (Rush et al., 2015), and NewsRoom (Grusky et al., 2018) are used as our external summary datasets for joint pretraining. All the three datasets are news articles or headlines with summaries from various news publications. We combine these datasets and the total training set consists of 5.6M samples.

4.2 Comparison Methods

For comparison, we select various baseline systems from previous literatures: the basic baseline **Longest-3** (Gliwa et al., 2019), which selects the longest three utterances as a summary; Classic seq2seq models, including **Seq2Seq+Attention** (Rush et al., 2015), **Transformer** (Vaswani et al., 2017), and **PGNet** (See et al., 2017); A pipeline method **FastRL** (Chen and Bansal, 2018) and its variant **FastRL Enhanced** (Gliwa et al., 2019), which first extracts salient sentences and then refines them; Convolution-based methods **LightConv** (Wu et al., 2019) and **DynamicConv** (Wu et al., 2019); Methods based on graph neural networks, including **D-HGN** (Feng et al., 2020) and **TGDGA** (Zhao et al., 2020); A seq2seq model **BERT+TRF** (Liu and Lapata, 2019) that is equipped with pretrained LMs.

4.3 Implementation Details

At the pretraining stage, we mix up the datasets from multiple sources and keep dialogues, short texts, and news summaries in a percentage of 1:1:1. The total data points are around 15M. Since DAMS consists of Transformer encoders and decoders, it can be easily combined with pretrained LMs. Here, we use BERT (Devlin et al., 2019) as the utterance/sentence encoder $\mathrm{TF}_{\theta_e^d}$ and use a separate optimization strategy (Liu and Lapata, 2019) to alleviate the mismatch between BERT and other randomly initialized parameters. We apply Adam (Kingma and Ba, 2015) (β_1=0.9, β_2=0.999) with learning rate 1e-3 for BERT and 1e-2 for other parameters. All transformer blocks except BERT have 6 layers, 8 heads, 768 hidden units, and the hidden size for all feed-forward layers is 2048. Loss coefficient α is selected from $\{0.01, 0.05, 0.1, 0.5\}$ to control adversarial signals, and we empirically find that $\alpha = 0.1$ achieves the best performance on the validation set. The model is pretrained for 250,000 steps with 10,000 warm-up steps on 2 GeForce RTX 3090 GPUs. At the fine-tuning stage, we use the last pretraining checkpoint for fine-tuning on the SAMSum dataset. We continue to train the model for 50,000 steps with 1,000 warm-up steps using Adam (β_1 =0.9, β_2=0.999, learning rate=1e-3). During the inference time, summaries are decoded in a beam size of 3. The minimal summary length is set to 15 for SAMSum and 100 for ADSC, respectively. Checkpoints are saved and evaluated on the validation set every 2,000 steps. The best checkpoint trained on SAMSum is directly evaluated on ADSC to perform zero-shot testing.

5 Results and Analysis

In this section, we show the main results of DAMS against other baselines for dialogue summarization, and probe the effectiveness of DAMS by explanatory experiments in various aspects.

[3] Following Feng et al. (2020), we train the model using SAMSum corpus and perform zero-shot testing on ADSC.

[4] https://github.com/nouhadziri/THRED

[5] Here, we use truncated sentence sequences in BookCorpus because we did not find other suitable corpora like MSCOCO. A real daily-life corpus with short-text summaries could be better for summary decoder pretraining.

Model	+News	RG-1	RG-2	RG-L
Longest-3	-	32.46	10.27	29.92
Seq2Seq+Att	-	29.35	15.90	28.16
Transformer	-	37.27	18.44	32.73
PGNet	-	40.08	15.28	36.63
FastRL	-	40.96	17.18	39.05
FastRL Enhanced	-	41.95	18.06	39.23
D-HGN	-	42.03	18.07	39.56
TGDGA	-	43.11	19.15	40.49
BERT+TRF	-	39.90	17.01	39.12
LightConv	✓	40.29	17.28	36.81
DynamicConv	✓	41.07	17.11	37.27
Transformer	✓	42.37	18.44	39.27
PGNet	✓	37.27	14.42	34.36
FastRL	✓	41.03	16.93	39.05
FastRL Enhanced	✓	41.87	17.47	39.53
BERT+TRF	✓	42.37	17.59	40.73
DAMS (w/o pretrain)	-	39.07	14.59	38.06
DAMS	✓	**44.38**	**19.98**	**43.40**

Table 2: Results of ROUGE-1/2/L on the SAMSum corpus. **+News** means whether the approach exploits external news summary data or not.

Model	RG-1	RG-2	RG-L
PGNet	28.95	5.34	22.41
Transformer	27.13	5.30	20.59
FastRL Enhanced	30.00	4.87	22.27
BERT+TRF*	28.13	4.63	27.17
DAMS (w/o pretrain)	28.17	5.11	27.09
DAMS*	**31.29**	**5.53**	**30.14**

Table 3: Results of zero-shot testing on ADSC. Models marked with * use external news summary data.

Methods	Informativeness	Fluency
PGNet	-0.128	-0.246
Transformer	-0.210	-0.119
FastRL Enhanced	-0.103	-0.052
BERT+TRF*	-0.037	0.091
DAMS*	**0.088**	**0.102**
Gold	0.390	0.224

Table 4: Human evaluation with model ranking results. Models with * utilize external news summary data.

5.1 Automatic Evaluation

Table 2 and Table 3 show the results of automatic evaluation on the SAMSum and ADSC dataset. We evaluate summary quality using ROUGE F1 (Lin, 2004), including the unigram and bigram overlap (ROUGE-1, ROUGE-2) between system outputs and gold summaries, and the longest common subsequence (ROUGE-L). Some results are from the reported scores in previous literatures (Gliwa et al., 2019; Feng et al., 2020; Zhao et al., 2020).

In Table 2, all baseline methods are categorized into two groups. The first group includes models that are directly trained on the SAMSum corpus, and methods in the second group benefit from external news summary data[6]. DAMS with full training data outperforms all baseline methods and is significantly different from BERT+TRF (+news) with $p < 0.05$, which probes the superiority of the multi-source pretraining strategy for dialogue summarization against the general exploitation of news summary data. Without news data, DAMS might be inferior to seq2seq models like PGNet or BERT+TRF, because these models use word-level attentions or copy mechanisms, while DAMS focuses on sentence/utterance representations for domain-agnostic representation learning. We also observe that the inclusion of news summary data does not necessarily mean a better ROUGE score (PGNet, FastRL). One possible explanation is that these models learn domain-specific features and have difficulty adapting to the dialogue domain. By contrast, with news summary data, the performance of DAMS increases a lot, which validates that our method can successfully capture useful information from external corpora. Furthermore, we directly test models on the ADSC dataset to verify whether they can generalize well to a new scenario. From Table 3 we observe that DAMS performs best, indicating that our multi-source pretraining strategy enables well-pretrained parameters for the downstream dialogue summarization, which makes the model easier to adapt to other dialogue scenarios.

5.2 Human Evaluation

Following Narayan et al. (2018), we randomly sample 100 examples in the test set of SAMSum for human evaluation. Three volunteers are invited to compare summaries produced from 6 systems (including the gold summary). Given a dialogue and two summaries from two out of six systems, each volunteer should decide which summary is better on two dimensions: **informativeness** (which summary captures more important information?) and **fluency** (which summary is more fluent?). We collect judgments from three volunteers for each comparison to minimize the inter-human noise.

Table 4 shows the system ranking results. Each score is calculated as the percentage of times the system is selected as best minus the percentage of

[6] For BERT+TRF, we pretrain it on our constructed news dataset. The other results are from Gliwa et al. (2019), where models are trained on the train set of CNN/DailyMail joined together with the train set of SAMSum, and evaluated on the SAMSum test set.

Methods	RG-1	RG-2	RG-L
DAMS	44.38	19.98	43.40
(w/o) D_e	42.29	18.33	41.28
(w/o) D_g	42.83	18.48	41.77
(w/o) D_e+D_g	43.89	18.52	42.09
(w/o) Dial.	42.89	18.17	41.60
(w/o) Short	43.01	18.65	41.71
(w/o) Summ.	43.37	17.98	41.65

Table 5: Ablation study of adversarial learning and multi-source pretraining. D_e, D_g are two critics. **Dial.**, **Short**, and **Summ.** denote corpora of dialogues, short texts, and news summaries, respectively.

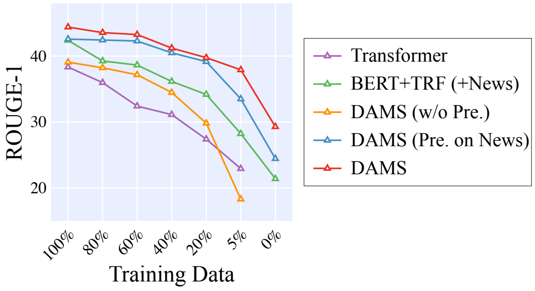


Figure 2: Model performance in low-resource settings.

times it is chosen as worst, ranging from -1 (worst) to 1 (best). **Gold** unsurprisingly ranks best. For informativeness, volunteers exhibit more preference to DAMS. For fluency, models with pretraining (DAMS / BERT+TRF) produce more acceptable summaries. We carry out pairwise comparisons between systems (using a binomial two-tailed test; p <0.05). In terms of informativeness, DAMS is significantly different from all other systems. For fluency, pretrain-based systems significantly differ from other systems, and BERT+TRF is not significantly different from DAMS.

5.3 Analysis and Discussion

We also perform qualitative analysis and discuss the effect of multi-source pretraining and adversarial learning with the following experiments.

Ablation Study. Table 5 shows the results of DAMS with different settings of adversarial critics and multi-source pretraining. We can see that the system suffers a performance degradation without the critic. It indicates that a domain-invariant representation is beneficial for downstream dialogue summarization. When any kind of external corpora is removed, the results drop a lot, which validates the effectiveness of multi-source pretraining.

Performance in Low-Resource Settings. To analyze model performances in low-resource settings, we gradually reduce the number of training examples in the SAMSum corpus by halving the training set. We report the results of DAMS and two baseline methods (Transformer and BERT+TRF) with different percentages of training data in Figure 2. We also report the performance of two variants of DAMS, without pretraining and only pretrained on the news summary data. Figure 2 shows a performance decline trend when the training data decreases continuously. We observe that with only limited SAMSum training data (40% / 20%), DAMS still achieves competitive results, while BERT+TRF (+News) suffers from a serious performance degradation. It indicates that DAMS has a promising ability of adapting news summaries to dialogue scenarios. Notably, using only 20% of the training data, DAMS achieves a competitive performance against Transformer and DAMS (w/o Pre.) that use the full training data, which proves the effectiveness of exploiting external corpora. When the training set is cut to 5% or even in a zero-shot setting, DAMS with multi-source pretraining shows a superior performance against all the other systems, including its variant DAMS (Pre. on News). It validates that our multi-source pretraining strategy is more applicable to dialogue summarization in a low-resource setting.

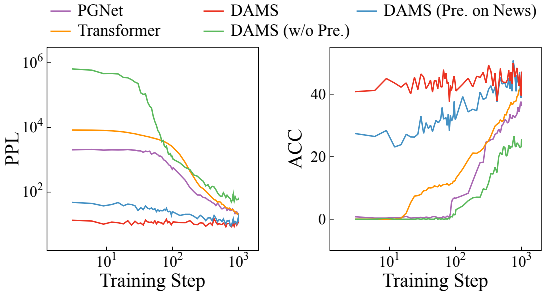
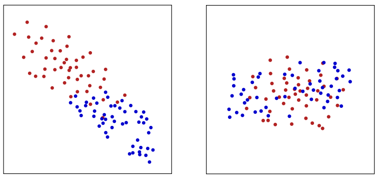


Figure 3: Fine-tuning logs of different models on the SAMSum dataset. PPL and ACC represent perplexity and word accuracy, respectively.

Convergence Rate. In Figure 3, we demonstrate the fine-tuning logs of different models on the SAMSum dataset. The left figure shows the perplexity and the right figure shows the average word accuracy. Unsurprisingly, models that benefit from pretraining have better initialized parameters, leading to faster convergence. Equipped with the multi-source pretraining strategy, DAMS can perform better and even achieve a 40% rate of word accuracy at the beginning of fine-tuning.



(a) w/o critic (b) w. critic

Figure 4: 2-D visualizations of representations in the dialogue and news domain.

Domain-Agnostic Representations. To verify the effectiveness of our adversarial strategy that can learn domain-agnostic features, we visualize the latent space of representations in 2-D using t-SNE (Van der Maaten and Hinton, 2008), with and without the critic. In Figure 4(a) where there is no critic, representations indeed show two separate clusters, while in Figure 4(b), hidden vectors with adversarial signals are effectively merged into one region, resulting in domain-agnostic representations. It encourages the summarizer to focus on content rather than domain-specific attributes for better generalization from other domains to the dialogue domain.

Case Study. Table 6 shows the system outputs of an exemplar dialogue. Texts with red color represent salient information in the dialogue, which is reflected in the gold summary. From the table we can see that DAMS can generate a summary that is more fluent and informative, which successfully captures critical information such as ’raining’ and ’half an hour’, composing a coherent discourse.

6 Conclusion and Future Work

In this paper, we propose a domain-agnostic multi-source pretraining paradigm for low-resource dialogue summarization, which exploits external large-scale corpora from multiple sources to facilitate dialogue modeling, summary language modeling, and abstractive summarization. The pretraining is conducted with adversarial signals to learn domain-agnostic summarization. The experimental results verify the effectiveness and generalization of our method in low-resource settings. Future directions are exploring how to keep the token-level cross attention in the multi-source pretraining strategy. In this way, we could adopt the strategy in the models with universal transformer architectures, e.g., BART, to benefit from large-scale pretraining language models.

Dialogue	Val : it’s raining! Candy: I know, just started... Val : r we going? we will be wet Candy: maybe wait a little? see if stops Val : ok. let’s wait half h and than see Candy: god idea, I call u then Val : great :)
Gold	It’s raining, so Val and Candy will wait half an hour before they go.
PGNet	Val is learning to meet Val and Val will see a little.
TRF	Val and Val don’t have any news. Val will call him because they got lost.
DAMS (w/o Pre.)	Candy and Val are going to meet. Val will call Candy instead.
BERT* +TRF	Val and Candy are going for a little, but they need to wait half an hour.
DAMS*	Val and Candy are going to wait half an hour to see if it’s raining.

Table 6: System outputs of a dialogue example from the SAMSum test set. Systems marked with * utilize external news summary data.

Acknowledgements

The authors wish to thank the anonymous reviewers for their helpful comments. This work was partially funded by China National Key R&D Program (No. 2017YFB1002104), National Natural Science Foundation of China (No. 61976056, 62076069), Shanghai Municipal Science and Technology Major Project (No.2021SHZDZX0103).

References

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686, Melbourne, Australia. Association for Computational Linguistics.

Eric Chu and Peter J. Liu. 2019. Meansum: A neural model for unsupervised multi-document abstractive summarization. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 1223–1232. PMLR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar Zaiane. 2019. Augmenting neural response generation with context-aware topical attention. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 18–31, Florence, Italy. Association for Computational Linguistics.

Xiachong Feng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2020. Incorporating commonsense knowledge into abstractive dialogue summarization via heterogeneous graph networks. *arXiv preprint arXiv:2010.10044*.

Yaroslav Ganin and Victor S. Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1180–1189. JMLR.org.

Dan Gillick, Korbinian Riedhammer, Benoit Favre, and Dilek Hakkani-Tur. 2009. A global optimization framework for meeting summarization. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4769–4772. IEEE.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Iryna Gurevych and Michael Strube. 2004. Semantic similarity applied to spoken dialogue summarization. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 764–770, Geneva, Switzerland. COLING.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1693–1701.

Xinyu Hua and Lu Wang. 2017. A pilot study of domain adaptation effect for neural abstractive summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 100–106, Copenhagen, Denmark. Association for Computational Linguistics.

Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2020. Dr. summarize: Global summarization of medical dialogue by exploiting local structures. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3755–3763, Online. Association for Computational Linguistics.

D Kingma and J Ba. 2015. Adam: A method for stochastic optimization in: Proceedings of the 3rd international conference for learning representations. *San Diego*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Chunyi Liu, Peng Wang, Jiang Xu, Zang Li, and Jieping Ye. 2019. Automatic dialogue summary generation for customer service. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pages 1957–1965. ACM.

Linqing Liu, Yao Lu, Min Yang, Qiang Qu, Jia Zhu, and Hongyan Li. 2018. Generative adversarial network for abstractive text summarization. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 8109–8110. AAAI Press.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Amita Misra, Pranav Anand, Jean E. Fox Tree, and Marilyn Walker. 2015. Using summarization to discover argument facets in online idealogical dialog. In *Proceedings of the 2015 Conference* *of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 430–440, Denver, Colorado. Association for Computational Linguistics.

Gabriel Murray and Giuseppe Carenini. 2008. Summarizing spoken and written conversations. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 773–782, Honolulu, Hawaii. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Xingchao Peng, Zijun Huang, Ximeng Sun, and Kate Saenko. 2019. Domain agnostic learning with disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5102–5112. PMLR.

Owen Rambow, Lokesh Shrestha, John Chen, and Christy Laurdisen. 2004. Summarizing email threads. In *Proceedings of HLT-NAACL 2004: Short Papers*, pages 105–108, Boston, Massachusetts, USA. Association for Computational Linguistics.

Banafsheh Rekabdar, Christos Mousas, and Bidyut Gupta. 2019. Generative adversarial network with policy gradient for text summarization. In *2019 IEEE 13th international conference on semantic computing (ICSC)*, pages 204–207. IEEE.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Oana Sandu, Giuseppe Carenini, Gabriel Murray, and Raymond Ng. 2010. Domain adaptation to summarize human conversations. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 16–22, Uppsala, Sweden. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Guokan Shang, Wensi Ding, Zekun Zhang, Antoine Tixier, Polykarpos Meladianos, Michalis Vazirgiannis, and Jean-Pierre Lorré. 2018. Unsupervised abstractive meeting summarization with multi-sentence compression and budgeted submodular maximization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 664–674, Melbourne, Australia. Association for Computational Linguistics.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, volume 307 of *ACM International Conference Proceeding Series*, pages 1096–1103. ACM.

Danqing Wang, Pengfei Liu, Ming Zhong, Jie Fu, Xipeng Qiu, and Xuanjing Huang. 2019. Exploring domain shift in extractive text summarization. *arXiv preprint arXiv:1908.11664*.

Lu Wang and Claire Cardie. 2013. Domain-independent abstract generation for focused meeting summarization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1395–1405, Sofia, Bulgaria. Association for Computational Linguistics.

Felix Wu, Angela Fan, Alexei Baevski, Yann N. Dauphin, and Michael Auli. 2019. Pay less attention with lightweight and dynamic convolutions. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Ziyi Yang, Chenguang Zhu, Robert Gmyr, Michael Zeng, Xuedong Huang, and Eric Darve. 2020. TED: A pretrained unsupervised summarization model with theme modeling and denoising. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1865–1874, Online. Association for Computational Linguistics.

Tiezheng Yu, Zihan Liu, and Pascale Fung. 2021. AdaptSum: Towards low-resource domain adaptation for abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, pages 5892–5904, Online. Association for Computational Linguistics.

Klaus Zechner. 2001. Automatic generation of concise summaries of spoken dialogues in unrestricted domains. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 199–207.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Xingxing Zhang, Furu Wei, and Ming Zhou. 2019. HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5059–5069, Florence, Italy. Association for Computational Linguistics.

Lulu Zhao, Weiran Xu, and Jun Guo. 2020. Improving abstractive dialogue summarization with graph structures and topic words. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 437–449, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online. Association for Computational Linguistics.

Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang. 2020. A hierarchical network for abstractive meeting summarization with cross-domain pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 194–203, Online. Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 19–27. IEEE Computer Society.

Yicheng Zou, Jun Lin, Lujun Zhao, Yangyang Kang, Zhuoren Jiang, Changlong Sun, Qi Zhang, Xuanjing Huang, and Xiaozhong Liu. 2021a. Unsupervised summarization for chat logs with topic-oriented ranking and context-aware auto-encoders. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14674–14682.

Yicheng Zou, Lujun Zhao, Yangyang Kang, Jun Lin, Minlong Peng, Zhuoren Jiang, Changlong Sun, Qi Zhang, Xuanjing Huang, and Xiaozhong Liu. 2021b. Topic-oriented spoken dialogue summarization for customer service with saliency-aware topic modeling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14665–14673.